Simultaneous Confidence Regions for the Multivariate Effective Dose

HANNA JANKOWSKI
hkj@mathstat.yorku.ca

Department of Mathematics and Statistics
York University

XIANG JI
jixiang@yorku.ca

Description of Problem

The effective dose (a.k.a. lethal/toxic dose) is the set of covariates which yield a specific response. Let $\mathcal{D} \subset \mathbb{R}^d$ denote the compact domain of the covariates. We assume the a binary response Y is observed and we are interested in sets of the form

$$\mathrm{ED}^+_{100p} = \{x \in \mathcal{D} : E[Y|X = x] \geq p\} \qquad \mathrm{ED}_{100p} = \{x \in \mathcal{D} : E[Y|X = x] = p\}$$

where $x = \{x_1, \ldots, x_d\}$ denotes the observed covariates. We assume that the data can be modelled as a logistic regression with

$$\log\left(\frac{E[Y|X = x]}{1 - E[Y|X = x]}\right) = \beta_0 x_0^* + \beta_1 x_1^* + \ldots + \beta_k x_k^* = \beta^T x^*,$$

where $x^*(x) : \mathbb{R}^d \mapsto \mathbb{R}^{k+1}$, with $k + 1 \geq d$, denotes a smooth function of the covariates. Let $\eta(u) = \log(u/(1-u))$, an increasing function of u. Therefore, we may re-write

$$\mathrm{ED}^+_{100p} = \left\{x \in \mathcal{D} : \beta^T x^* \geq \eta(p)\right\} \qquad \mathrm{ED}_{100p} = \left\{x \in \mathcal{D} : \beta^T x^* = \eta(p)\right\}.$$

We estimate these using the plug-in estimators

$$\widehat{\mathrm{ED}}^+_{100p} = \left\{x \in \mathcal{D} : \widehat{\beta}_n^T x^* \geq \eta(p)\right\} \qquad \widehat{\mathrm{ED}}_{100p} = \left\{x \in \mathcal{D} : \widehat{\beta}_n^T x^* = \eta(p)\right\},$$

where $\widehat{\beta}_n$ is the well-known maximum likelihood estimator of β. Our goal is to develop a simultaneous confidence region for ED^+_{100p} and ED_{100p}.

Calculating the Confidence Regions

Below, we explain how to calculate the $100(1-\alpha)\%$ confidence region for ED^+_{100p}. A $100(1-\alpha)\%$ confidence region for ED_{100p} is taken as the intersection of the $100(1-\alpha/2)\%$ confidence region for ED^+_{100p} and the $100(1-\alpha/2)\%$ confidence region for $\overline{(\mathrm{ED}^+_{100p})^c}$.

(1) INVERTING SCHEFFE'S BOUNDS

A conservative, and currently only, method developed by Carter et al. (1986) is based on inverting Scheffé's upper bound. Recall that $\lim_n \sqrt{n}(\widehat{\beta}_n - \beta) \sim N_{k+1}(0, \Sigma)$. Then the $100(1-\alpha)\%$ confidence region for ED^+_p is

$$\mathrm{CR}^+_{\mathrm{SCH},100(1-\alpha)} = \left\{x : \widehat{\beta}_n^T x^* \geq \eta(p) - \sqrt{\chi^2_\alpha(k+1)\, x^{*T}\widehat{\Sigma} x^*/n}\right\}.$$

where $\widehat{\Sigma}$ is the usual estimate of Σ. This approach was studied extensively by Li et al. (2008a) in the linear setting.

(2) VOROBE'EV QUANTILE

Let $\mathbf{A}$ be a random closed set (RCS) and define $T(K) = P(\mathbf{A} \cap K \neq \emptyset)$. As defined in Molchanov (1990), the p–quantile of $\mathbf{A}$ is $M_p = \cup\{K \in \mathcal{M} : T(K) < p\}$, and we choose $\mathcal{M} = \left\{\{x\}; x \in \mathbb{R}^d\right\}$. Then the $100(1-\alpha)\%$ confidence region for ED^+_p is

$$\mathrm{CR}^+_{\mathrm{Q},100(1-\alpha)} = M^c_\alpha,$$

where M_α denotes the quantile of $\widehat{\mathrm{ED}}^+_p$. This is estimated empirically via re-sampling. Heuristically, the confidence region is the collection of points $\{x\}$ such that each falls inside $\widehat{\mathrm{ED}}^+_p$ at least $100(1-\alpha)\%$ of the time.

(3) NEW DEFINITION OF RCS QUANTILE

Let $\mathcal{A}$ be a collection of sets such that $\mathbf{A} \in \mathcal{A}$ almost surely. Fix a set K, and define the random variable $\xi = \rho_H(\mathbf{A}, K)$. Let q^ξ_p denote the p–quantile of ξ. The quantile of $\mathbf{A}$ is defined as $M^H_p = \cup\left\{A \in \mathcal{A} : \rho_H(A, K) \leq q^\xi_p\right\}$, where ρ_H denotes the Hausdorff distance. The $100(1-\alpha)\%$ confidence region for ED^+_p is then

$$\mathrm{CR}^+_{100(1-\alpha)} = M^H_{1-\alpha},$$

where $M^H_{1-\alpha}$ is the quantile of $\widehat{\mathrm{ED}}_p$. Draw $A_1, \ldots, A_n$ samples of ED^+_p and calculate $K_n = \cap_{i=1}^n A_i$ with $\xi_{i,n} = \rho_H(A_i, K_n)$. Then the estimate of the confidence region is

$$\cup\left\{A_i, i = 1, \ldots, n : \rho_H(A_i, K_n) \leq \widehat{q}^\xi_{n,1-\alpha}\right\},$$

where $\widehat{q}^\xi_{n,p}$ is the quantile estimate from the observations of $\xi_{i,n}$.

Results

We consider the following four models.

model	true value of $\beta^T x^*$	domain
linear	$-6 + 6x_1 + 6x_2$	$[0,1]^2$
interaction	$-6 + 6x_1 + 6x_2 - 3x_1x_2$	$[0,1]^2$
quadratic	$-6 + 6x_1 + 6x_2 + 10x_1^2 + 3x_1x_2 + x_2^2$	$[0,1]^2$
log term	$-10 + 6\log x_1 + 6x_2$	$[1,2]^2$

To compare our methods, we ran Monte Carlo simulations to find the empirical coverage probabilities. We simulate ED_{100p} values under the asymptotic distribution of $\widehat{\beta}_n$. The results are shown in the tables below.

	linear			interaction			quadratic			quadratic (+)			log term		
p	**(1)**	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**
0.1	.99	.89	.97	1.00	.89	.97	1.00	.87	.99	1.00	.87	.99	.99	.92	.98
0.5	.98	.86	**.94**	1.00	.79	**.94**	1.00	.80	.98	1.00	.81	.97	.99	.87	**.96**
0.9	.99	.90	.99	1.00	.91	1.00	1.00	.80	.97	1.00	.80	**.96**	.98	.88	**.96**

Table 1: Empirical coverage results for 95% confidence regions when simulating from the limiting distribution. The sample size is 360, except in the fourth column (quadratic (+)), where the sample size is 3600. Results not statistically different from 0.95 are shown in bold.

	linear			interaction			quadratic			quadratic (+)			log term		
p	**(1)**	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**	**(1)**	**(2)**	**(3)**
0.1	.18	.12	.17	.26	.16	.21	1.00	.15	.30	1.00	.03	.05	.10	.07	.10
0.5	.19	.13	.20	.27	.17	.27	1.00	.12	.47	1.00	.03	.05	.18	.12	.19
0.9	.18	.12	.19	.12	.09	.12	1.00	.24	.66	1.00	.04	.07	.26	.18	.28

Table 2: Median proportion of the domain covered by the confidence region. The simulations are the same as those shown in Table 1.

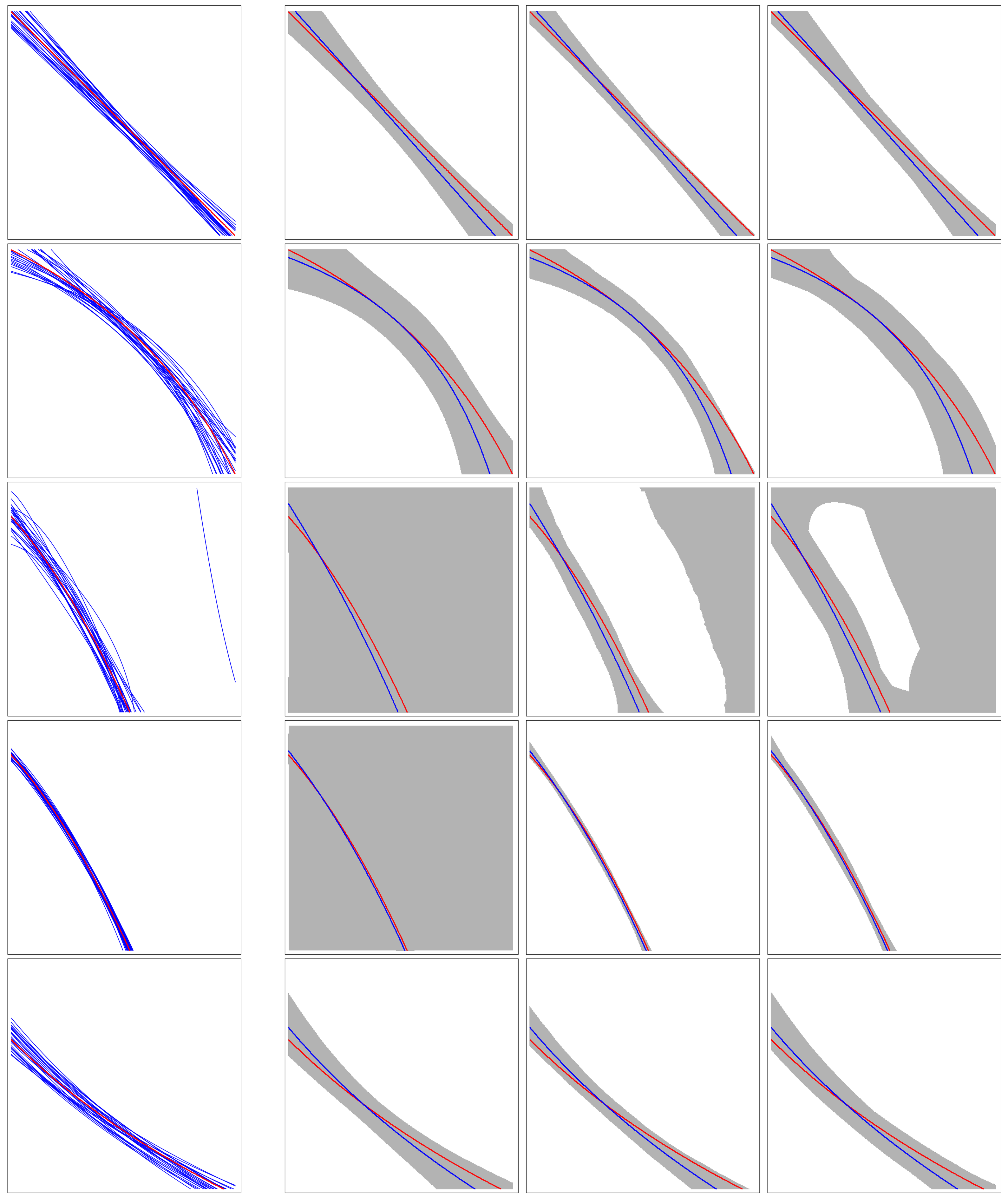

Figure 1: Examples of the different methods: In the leftmost column are sample observations of $\widehat{\mathrm{ED}}_{50}$ and the three columns on the right show the three regions **(1)** $\mathrm{CR}_{\mathrm{SCH},95}$, **(2)** $\mathrm{CR}_{\mathrm{Q},95}$, and **(3)** CR_{95} in light grey, from left to right. Observed values of $\widehat{\mathrm{ED}}_{50}$ are shown in blue, and the true set ED_{50} is shown in red. From top to bottom the models are linear ($n = 360$), interaction ($n = 360$), quadratic ($n = 360$), quadratic ($n = 3600$), and log term ($n = 360$).

References

Carter, W., Chinchilli, V., Wilson, J., Campbell, E., Kessler, F. and Carchman, R. (1986). An asymptotic confidence region for the ED100p from the logistic response surface for a combination of agents. *The American Statistician* **40** 124–128.

Li, J., Zhang, C. M., Nordheim, E. V. and Lehner, C. E. (2008a). Estimation and confidence regions of multi-dimensional effective dose. *Biometrical Journal* **50** 110–112.

Molchanov, I. (1990). Empirical estimation of distribution quantiles of random closed sets (Russian). *Teor. Veroyatnost. i Primen.* **35** 586–592. English translation in: *Theory of Probability and Its Applications*, 1990, **35**, pp.594-600.